If we have $k-1$ regressors then we have 2^{k-1} models, how do we choose the best?

All the models are evaluated using the following criteria.

1) Coefficient of determination (R^2)
2) Adjusted R^2
3) Mallow's Statistic C_p
4) Akaike's Information Criteria (AIC)
5) Bayesian Information Criteria (BIC)

1) R_p^2 : $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$

$$R_p^2 = \frac{SS_{Reg}(p)}{SS_T} = 1 - \frac{SS_{Res}(p)}{SS_T}$$ indep. of p!!!

R_p^2 is maximum when we have a full model

R_p^2 explains the percentage of total variability which is explained by the model from the response variable.

* Thumb Rule: Choose the model with the highest R^2

However, it does not consider newly added features and does not have p in the formula (No p in SLR)

2) Adjusted $R^2 = 1 - \frac{MS_{Res}}{MS_T}$

$$= 1 - \frac{SS_{Res}}{n-p} \cdot \frac{n-1}{SS_T}$$

$$= 1 - \left(\frac{n-1}{n-p}\right) R^2$$

For $n \to +\infty$, Adjusted $R^2 \approx R^2$, the higher the adjusted R^2, the better the model.

3) Mallow C_p : Measures the overall $MSE = Var + bias^2$ in the fitted model

$$C_p = MSE + \text{penalty}$$ → estimate of bias (cost) on model having extra param.

$$= \frac{SS_{Res}(p)}{MS_{Res}^{Full}} + (-n+2p)$$

$$C_K = \frac{SS_{Res}(Full)}{MS_{Res}^{Full}} + (-n+2K)$$

$$= (n-K) + (-n+2K)$$

$$= K$$

Lower C_p gives better model, when C_p is close to p gives the best model e.g :

$$S = \{X_1, X_2\}$$

M1 : $Y = \beta_0 + \beta_1 X_1$ → $C_p \approx 2$

M2 : $Y = \beta_0 + \beta_2 X_2$ → $C_p \approx 2$

M3 : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ → $C_p \approx 3$